Integration

(National Center for **Integrative** Biomedical Informatics)

- Many different sources of biomedical data
  - Lab experiments
  - Published literature
  - Public (and private/commercial) databases
- Many different tools to manipulate biomedical data

Software Integration

- A software resources repository is in place
  - simple web form for creators to register new software.
- Work with cross-NCBC group (SDIWG) to enhance functionality.
- Modularize software and string it into a workflow
  - Initial examples in use
  - More comprehensive vision
- Understand constraints on workflows to suggest legal compositions.
  - Aerospace Corporation

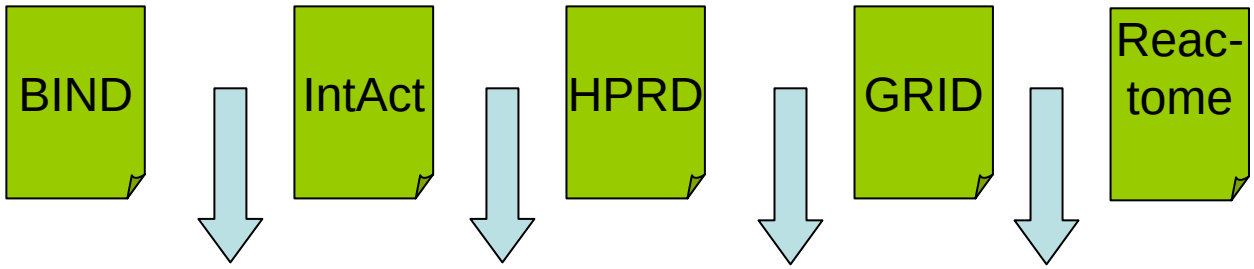
Data Integration

- Ontology mapping
  - FAM to MESH
  - GO to MESH
  - Standardization, through cross-NCBC effort.
- Database integration
  - MBI
  - MiMI

Motivation

- Copious amounts of protein data exist online
- Some of it is repeated across sources, some of it is contradictory between sources
- Experiments used to furnish data have varying levels of false positives and negatives
- Researchers must get pieces from disparate sources and piece them together manually, making judgments about the quality of each source as they work.
  - Users often require context for data that they see, e.g.
    - type of experiment used,
    - the organism, the tissue, etc.
  - Deep integration provides an opportunity to obtain and represent such relevant context, even if this requires consulting additional data sources.
- One possible solution is: MiMI

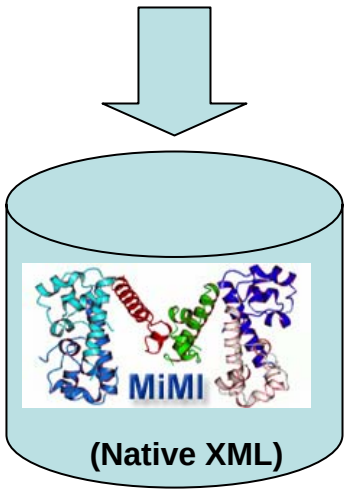
MiMI Data Sources, Deep Data Integration and Access

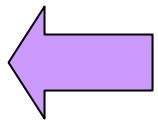


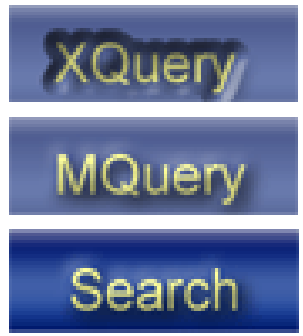
Deep Integration Techniques

Deep Integration Techniques:

- **Preserves provenance and conflicts**
- **Uniquely identify Proteins across multiple sources**
- **Integrates knowledge bases**







MiMI is accessible to both expert and novice users by providing:

- **Intelligent search & form-based queries**
- **XQuery**

What is MiMI ?

- Comprehensive database of molecular interactions.
- Merges data into unique objects
  - source data: BIND, DIP, GRID, IntAct, HPRD, the Center for Cancer Systems Biology at Harvard, the Max Delbrueck Center, and Reactome
  - Auxilliary data: InterPro, IPI, GO, OrganelleDB, Pfam, OrthoMCL, ProtoNet, miBLAST, SwissProt.

MiMI cont.

- Merged information generates:
  - 258,953 molecules
  - 408,980 interactions
  - 10,953 complexes
  - 8,589 pathways.
- Adopts XML technology for easy transformation between data sources; employs the native XML database TIMBER for efficient and effective retrieval

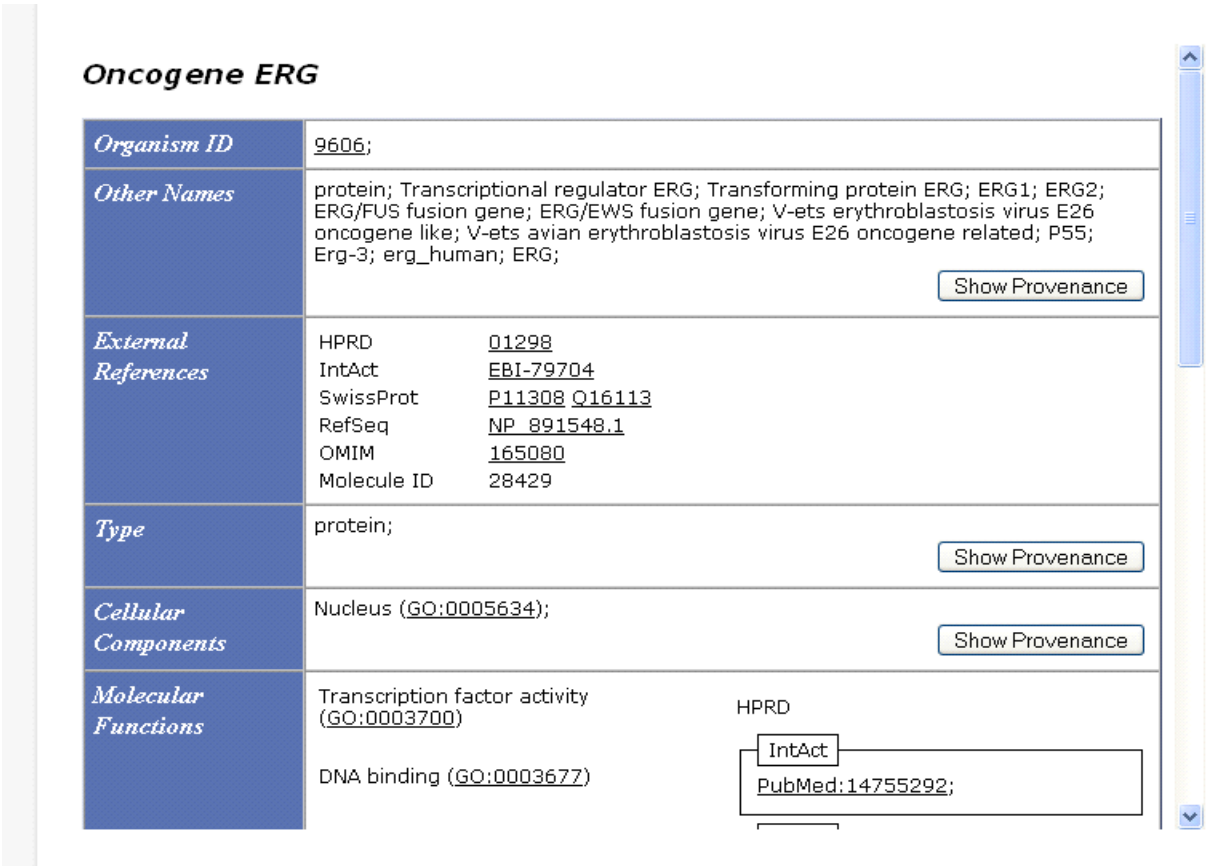
Research and Techniques in MiMI

- Timber (using TAX) http://blahblahblah
- Efficient XQuery parser
- Open Source Standards Based Software
- Deep integration
- Provenance Prototype
- Multiple querying options

Deep Integration

- Deep Integration means the original source data has been transformed, and is not directly available.
- Address, by maintaining Provenance
- Track how each piece of data is obtained.
    - Provides credit to data provider
    - Enhances user confidence in data.
    - Facilitates resolution of disagreement/error.
- Research on efficient storage of provenance.
- Allows scientists to query data based on provenance

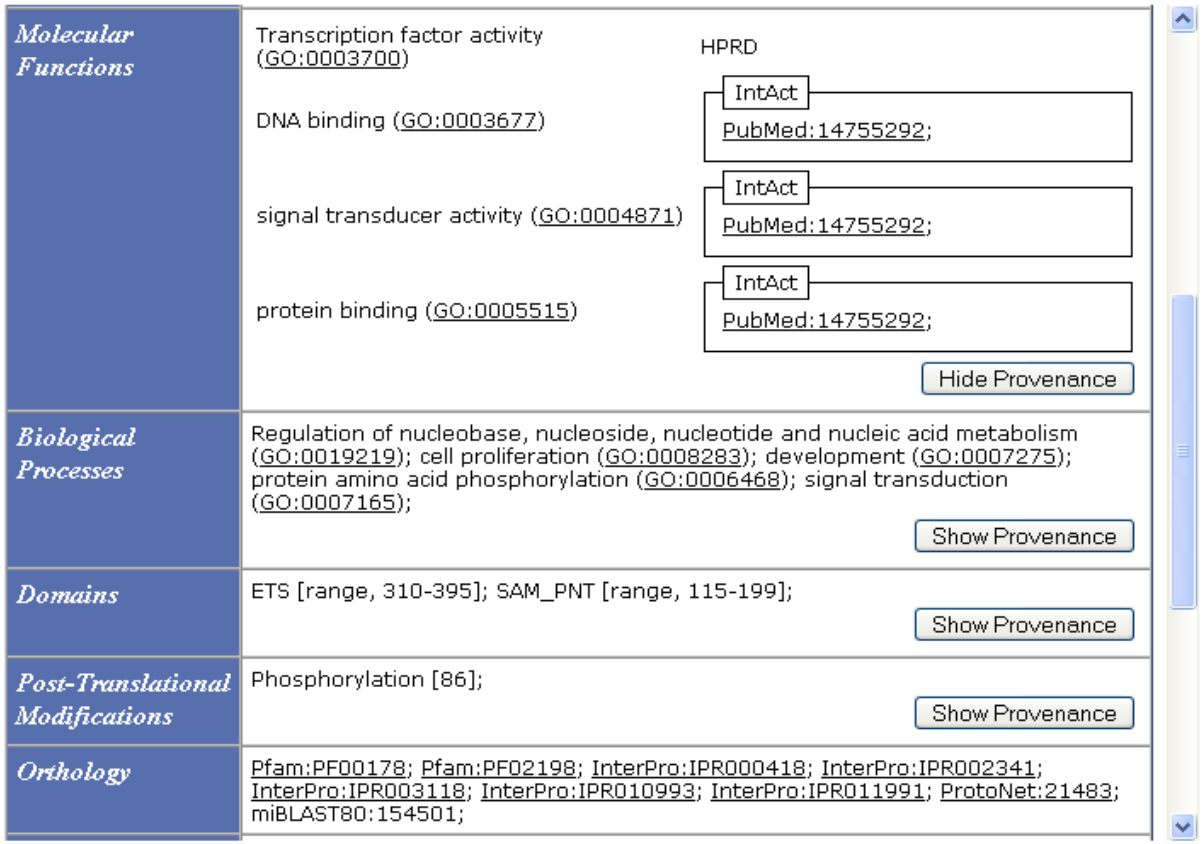
ERG (Prostate)



Oncogene ERG

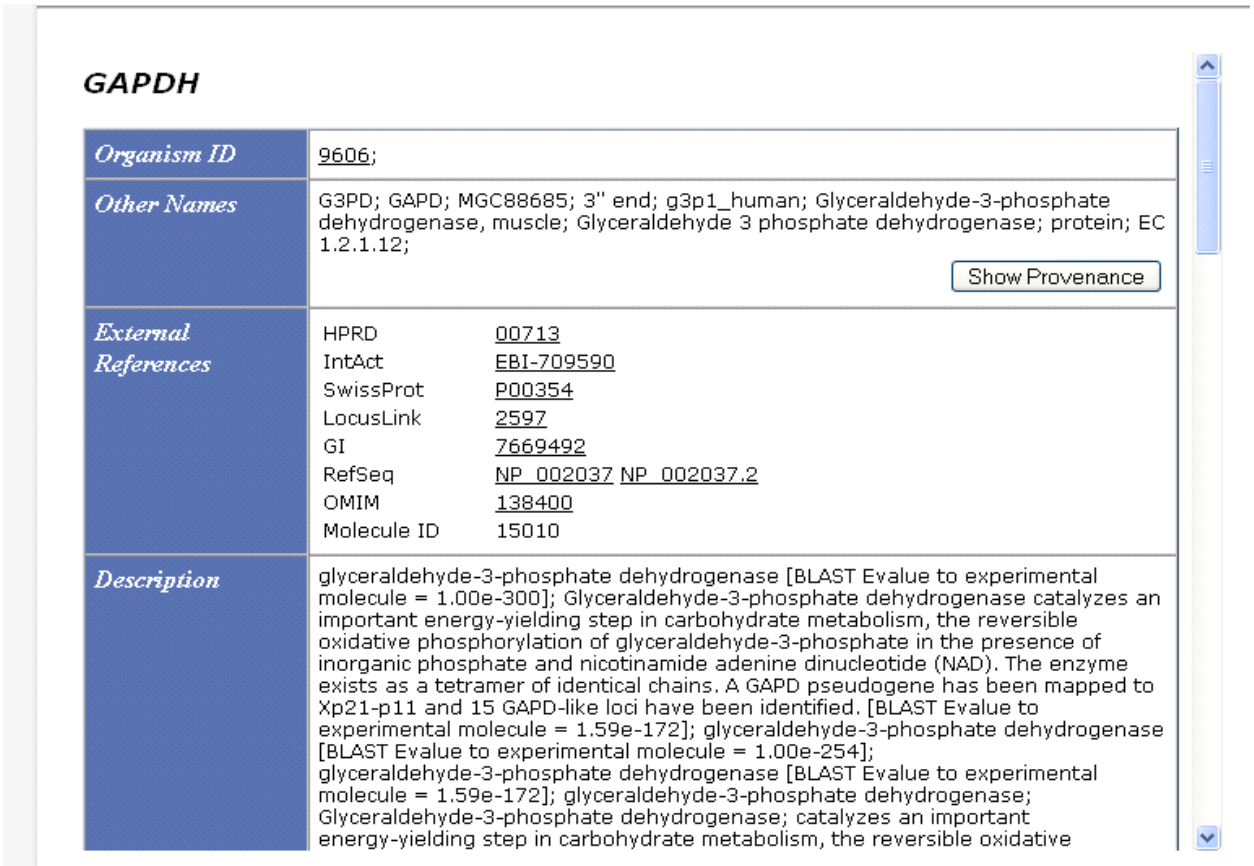
Organism ID	9606;
Other Names	protein; Transcriptional regulator ERG; Transforming protein ERG; ERG1; ERG2; ERG/FUS fusion gene; ERG/EWS fusion gene; V-ets erythroblastosis virus E26 oncogene like; V-ets avian erythroblastosis virus E26 oncogene related; P55; Erg-3; erg_human; ERG; Show Provenance
External References	HPRD 01298 IntAct EBI-79704 SwissProt P11308 Q16113 RefSeq NP_891548.1 OMIM 165080 Molecule ID 28429
Type	protein; Show Provenance
Cellular Components	Nucleus (GO:0005634); Show Provenance
Molecular Functions	Transcription factor activity (GO:0003700) DNA binding (GO:0003677) HPRD IntAct PubMed:14755292;

ERG Slide 2



Molecular Functions	Transcription factor activity (GO:0003700) — HPRD DNA binding (GO:0003677) — IntAct: PubMed:14755292; signal transducer activity (GO:0004871) — IntAct: PubMed:14755292; protein binding (GO:0005515) — IntAct: PubMed:14755292; Hide Provenance
Biological Processes	Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism (GO:0019219); cell proliferation (GO:0008283); development (GO:0007275); protein amino acid phosphorylation (GO:0006468); signal transduction (GO:0007165); Show Provenance
Domains	ETS [range, 310-395]; SAM_PNT [range, 115-199]; Show Provenance
Post-Translational Modifications	Phosphorylation [86]; Show Provenance
Orthology	Pfam:PF00178; Pfam:PF02198; InterPro:IPR000418; InterPro:IPR002341; InterPro:IPR003118; InterPro:IPR010993; InterPro:IPR011991; ProtoNet:21483; miBLAST80:154501;

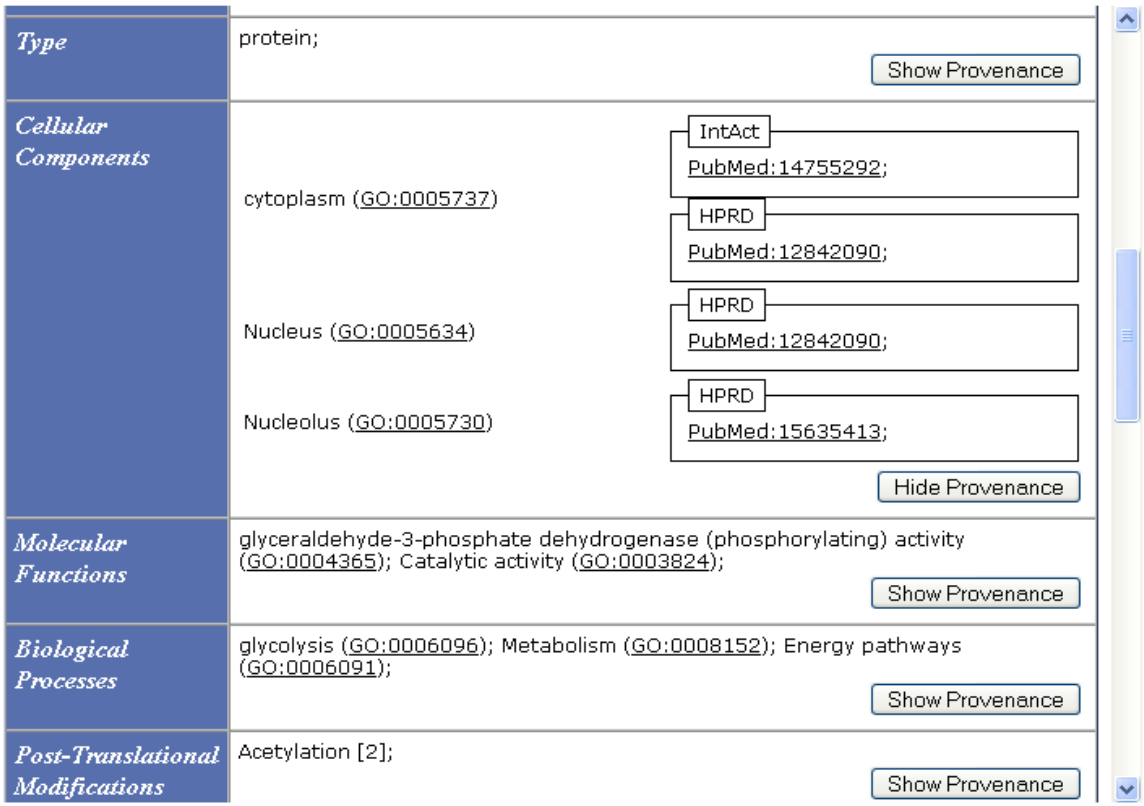
HNF4A (Diabetes)



GAPDH

Organism ID	9606;
Other Names	G3PD; GAPD; MGC88685; 3" end; g3p1_human; Glyceraldehyde-3-phosphate dehydrogenase, muscle; Glyceraldehyde 3 phosphate dehydrogenase; protein; EC 1.2.1.12; Show Provenance
External References	HPRD 00713 IntAct EBI-709590 SwissProt P00354 LocusLink 2597 GI 7669492 RefSeq NP_002037 NP_002037.2 OMIM 138400 Molecule ID 15010
Description	glyceraldehyde-3-phosphate dehydrogenase [BLAST Evalue to experimental molecule = 1.00e-300]; Glyceraldehyde-3-phosphate dehydrogenase catalyzes an important energy-yielding step in carbohydrate metabolism, the reversible oxidative phosphorylation of glyceraldehyde-3-phosphate in the presence of inorganic phosphate and nicotinamide adenine dinucleotide (NAD). The enzyme exists as a tetramer of identical chains. A GAPD pseudogene has been mapped to Xp21-p11 and 15 GAPD-like loci have been identified. [BLAST Evalue to experimental molecule = 1.59e-172]; glyceraldehyde-3-phosphate dehydrogenase [BLAST Evalue to experimental molecule = 1.00e-254]; glyceraldehyde-3-phosphate dehydrogenase [BLAST Evalue to experimental molecule = 1.59e-172]; glyceraldehyde-3-phosphate dehydrogenase; Glyceraldehyde-3-phosphate dehydrogenase; catalyzes an important energy-yielding step in carbohydrate metabolism, the reversible oxidative

HNF4A Slide 2

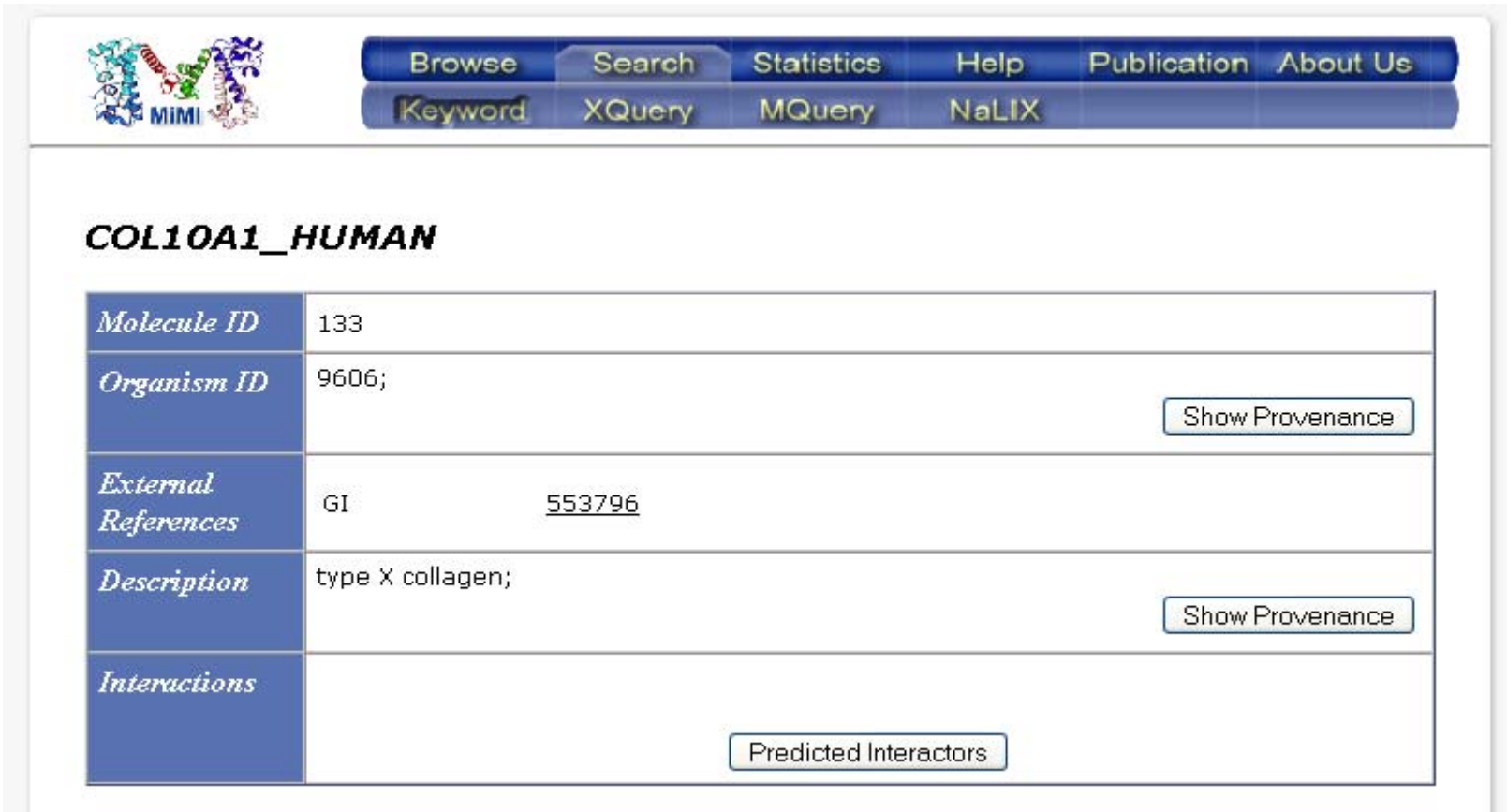


Type	protein;	Show Provenance
Cellular Components	cytoplasm (GO:0005737)	IntAct PubMed:14755292; HPRD PubMed:12842090;
	Nucleus (GO:0005634)	HPRD PubMed:12842090;
	Nucleolus (GO:0005730)	HPRD PubMed:15635413;
		Hide Provenance
Molecular Functions	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity (GO:0004365); Catalytic activity (GO:0003824);	Show Provenance
Biological Processes	glycolysis (GO:0006096); Metabolism (GO:0008152); Energy pathways (GO:0006091);	Show Provenance
Post-Translational Modifications	Acetylation [2];	Show Provenance

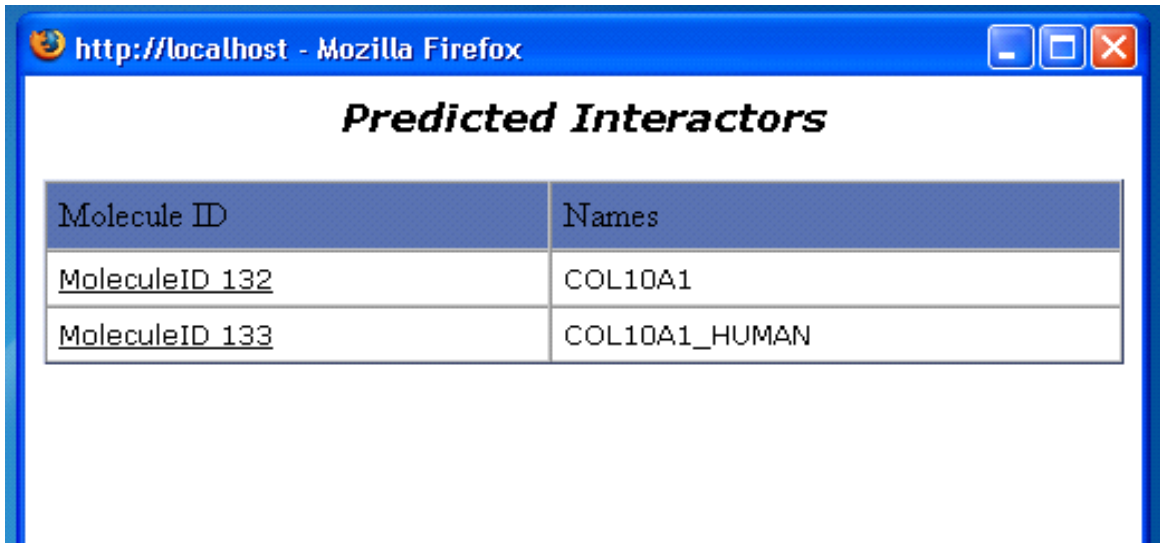
Interaction Predictions

- Interactions can also be predicted based on several methods:
  - *Structure-based prediction*: two proteins with potentially interacting domains are likely to interact with each other
  - *Homology-based prediction*: two proteins are likely to interact with each other if their respective homologues interact
  - Various other factors affect “interactibitlity”, in particular, expression profiles (both temporal and spatial).
- MiMI keeps track of both homologue and structural information
  - Currently, MiMI facilitates automatic interaction prediction based on homology.
  - In the future, predicted interactions will be subjected to verification based on expression profiles.

Predicted Interactors Step 1



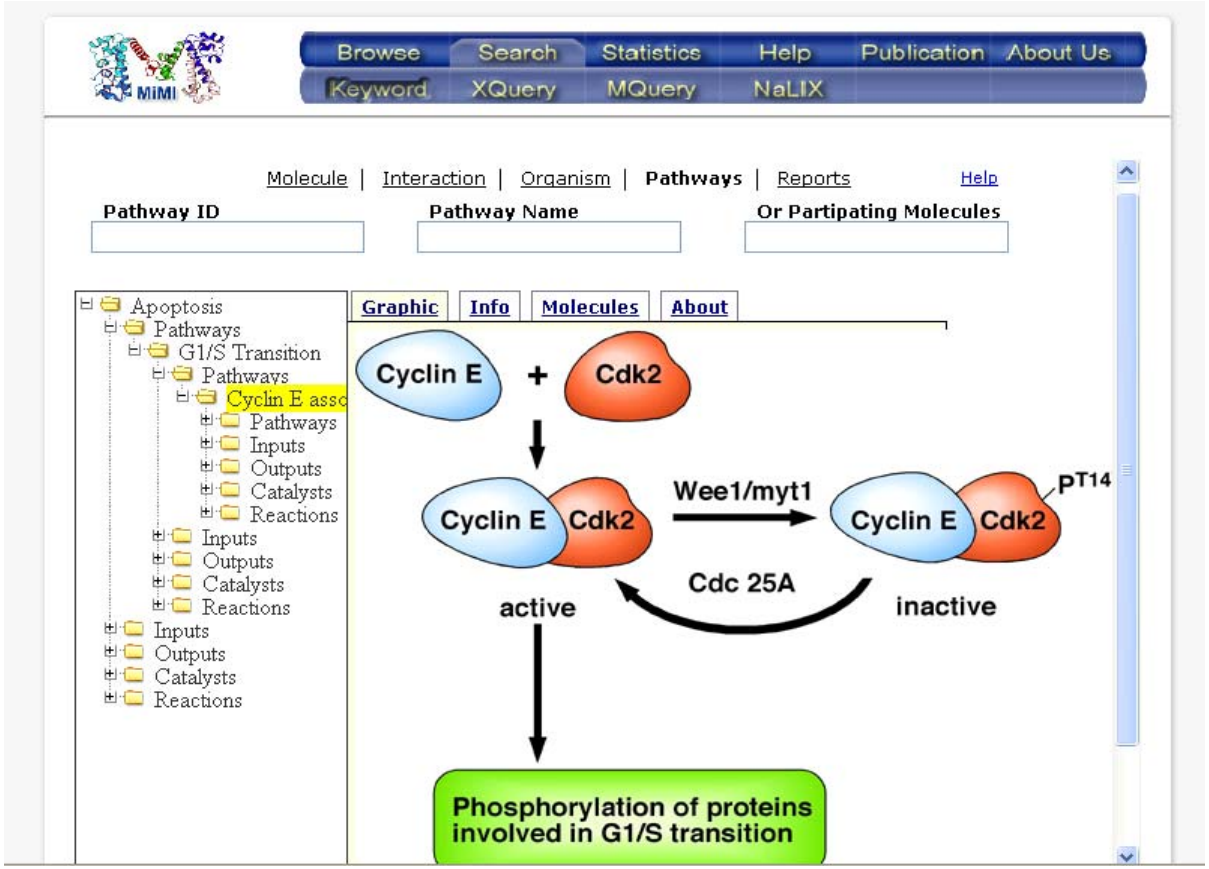
Predicted Interactors Step 2



Schema Rewriting and Simplificiation

- Biological schema are complex and non-uniform
- Transformation of data is a manual and error prone task
- Automated schema rewriting helps with
  - Understanding the schema
  - Transforming and integrating the data
- Future data integration into MiMI will be facilitated using methods we have developed

Pathways


MiMI
Browse
Search
Statistics
Help
Publication
About Us
Keyword
XQuery
MQuery
NaLIX
Molecule | Interaction | Organism | Pathways | Reports
Help
Pathway ID
Pathway Name
Or Partipating Molecules
Apoptosis
Pathways
G1/S Transition
Pathways
Cyclin E assc
Pathways
Inputs
Outputs
Catalysts
Reactions
Inputs
Outputs
Catalysts
Reactions
Inputs
Outputs
Catalysts
Reactions
Graphic
Info
Molecules
About
Cyclin E
+
Cdk2
Cyclin E
Cdk2
Wee1/myt1
Cyclin E
Cdk2
PT14
Cdc 25A
active
inactive
Phosphorylation of proteins involved in G1/S transition

Conclusions

- MiMI is available at
  - http://mimi.ctaalliance.org
- MiMI provides a new and useful tool to biologists that compliments and increases the value of current interaction databases:
  - Its deep integration feature provides a one-stop shop for biomedical scientists interested in comprehensive information about particular interactions or pathways.
  - Its provenance and probability tracking feature gives scientists more control over what information to trust.
  - Its flexible querying interfaces allows scientists to obtain information filtered and customized to their needs.